Examining tacit knowledges in assessing international postgraduate students

Examining Tacit Knowledges in Assessing International Postgraduate Students

Claudia Rosenhan[a]*, Farah Akbar[b] and Takuya Numajiri[c]

[a] *Moray House School of Education, University of Edinburgh, Edinburgh, UK;* [b] *Moray House School of Education, University of Edinburgh, Edinburgh, UK;* [c] *Moray House School of Education, University of Edinburgh, Edinburgh, UK.*

*Email: claudia.rosenhan@ed.ac.uk

Claudia Rosenhan, PhD, FHEA, works as a teaching fellow on a large PGT programme in Language Education.

Farah Akbar, PhD, works as a teaching fellow on a large PGT programme in Language Education.

Takuya Numajiri is a PhD candidate at Moray House.

Abstract

The study investigates student and scorer attitudes to a marking scheme used on a taught postgraduate programme, to examine whether level descriptors enhance students' and staff assessment literacy. The student cohort was surveyed at two time-points, with a response rate of 62% ($N = 99$) and 24% ($N = 39$) respectively. One focus group with four scorers was also conducted. Using exploratory factor analysis, we found that students were confident in their understanding of the descriptors, but also believed that markers draw on tacit knowledges. This concern was confirmed to an extent by the focus group. The findings question the usefulness of descriptors to foster assessment literacy, especially for international students, as they do not mitigate against tacit knowledge. Both data sets were small and therefore not generalizable. The findings are, however, indicative of recurring issues in academic assessment, in which international students struggle to attain the requisite understanding of quality necessary for their development as autonomous learners.

Keywords: postgraduate; assessment; assessment literacy; higher education

Introduction

It has been noted that students and scorers in higher education (HE) often suffer from low assessment literacy, i.e. they lack a full grasp of assessment principles and practices (Norton et al., 2013; Price et al., 2012). The way staff and students engage with a marking scheme can thus be considered a touchstone for examining and honing this skill (see ***Image 1*** for outline of the marking scheme and level descriptors on which this investigation is based). Level descriptors are part of the marking scheme, providing the quality definitions for the evaluative criteria at determined levels. In the context of this investigation, a taught postgraduate (PGT) programme with a high level (over 90%) of international students, the marking scheme is designed along a task-type rubric for a critical academic essay, to be used generically across a range of similar assignments (***Image 1***).

Image 1: Marking Scheme

	Knowledge and understanding of Concepts	Knowledge and Use of Literature	Critical Reflection on Theory and Practice	Application to Theory and Practice	Constructing Academic Discourse	Planning and Implementation of the Research/Investigation (used mainly for dissertations)
A (70 – 100%) Distinction						
B (60 – 69%) Merit	*E.G. The work demonstrates understanding of the concepts and theories relevant to the assignment task through outlining and reporting on established issues with a view to explanation. Judgement is used to establish relationships between the various relevant concepts and theories, but these are not evaluated or framed by different perspectives. There is a tendency to list the concepts or place them in the argument without further reflection.*	Level descriptors				
C (50 – 59%) Good pass						
D (40 – 49%) Pass at Diploma Level						
E (30 – 39%) Fail						
F (below 30%) Bad Fail		Not addressed in this investigation				

It is intended to help articulate consistent feedback in accordance with the marker's professional judgement, to guide student learning, and enable them to improve for subsequent assignments. However, if students cannot relate especially to the descriptors, or if scorers use them inconsistently, the assessment may fail to foster student learning. We wanted to investigate, therefore, students' attitudes towards the marking scheme in general, and the level descriptors in particular. By understanding students' attitudes, we would be able to infer whether students felt they were successfully engaging in a dialogue with their scorers via the level descriptors (Nicol, 2010), thus addressing their assessment literacy. At the same time, we were also interested in the attitude of the scorers, and how they utilise the marking scheme in their assessment of student work and the feedback they provide. Our investigation was, therefore, designed as a mixed-method research in which we investigated students' attitudes quantitatively through a questionnaire survey and factor analysis, and scorer perspectives qualitative through a focus group and thematic analysis.

This was considered relevant, because assessment and feedback regularly receives the

highest level of negative responses in postgraduate surveys. Students report, for example, a lack of clarity about what they are expected to achieve (PTES, 2017). To counteract these concerns, it is important to develop students' understanding of the quality of their work and hone their judging skills. Nicol et al. (2014) note the active role students must play in such processes. The descriptors on the programme had recently been rewritten, with the expressed intention of making criteria more explicit, and thus enhance their use as a medium for shared understanding and dialogue between scorers and students. It was, therefore, of interest to gauge student and staff attitudes on a postgraduate programme towards the level descriptors through a mixed-method research, and thus infer their level of active engagement. The research questions to investigate this phenomenon were:

- Do level descriptors assist a clear understanding of the criteria amongst students (thereby enhancing assessment literacy)?
- Do level descriptors assist a clear understanding of the criteria amongst groups of scorers on assignments (thereby enhancing scorer reliability)?

Literature Review

The topic of assessment literacy in higher education (HE) is salient, in that it goes to the heart of how value is being added for students to achieve their learning objectives. This is particularly relevant for international students, who frequently struggle with new academic cultures. Level descriptors could be considered a direct way of scaffolding international students' understanding of how to develop their learning, especially since the descriptors operate at course or programme level.

The debate on descriptors originated, however, in the link between criterion-based assessment of educational outcomes and quality assurance at national level. National quality codes demand accountability, explicitness and constructive alignment from assessment

processes (e.g. Quality Assurance Agency for Higher Education, 2013), hence quality in higher education is directly associated with assessment criteria and descriptors. Boud (2007) found that quality assurance aspects dominate the assessment policies of HE institutions, and exhaustive documentation of assessment processes and moderation practices are standard procedure for any given university. Standards knowledge is thus encoded and disseminated via the numerous artefacts, of which the marking scheme is but one example (Sadler, 2014).

Grainger et al. (2008) suggest that published criteria and descriptors serve as a strategy to address public scepticism about educational standards. The key idea is that standards are maintained through transparency and public accountability (Brown, 2010; Koh, 2011). If descriptors align with established performance criteria, they are more likely accepted publicly as a trustworthy indicator for quality. Criterion referencing further suggests that assessment in HE is an objective and robust analytical measurement, replacing a perceived arcane standards model that relied, as Stowell (2004) states, on undefined assumptions. This techno-rationalist paradigm, in line with the auditable outcomes-based ethos of HE (Hussey and Smith, 2002), is meant to cast a "veil of rigour" over what remains a fundamentally subjective assessment method of complex intellectual performances required by postgraduate students. Pre-set criteria, in any case, can only record a fuzzy signal of achievement and overlook performances that are not articulated. While descriptors are useful in recording the essence of a performance standard across different levels, they do, as for example Bloxham (2009) has found, not automatically guarantee good quality assessment practices. Overall, commentators find that faith in criterion-referencing is misplaced.

'Constructive Alignment' (Biggs and Tang, 2010) has, nevertheless, become, according to Hudson et al. (2017), the dominant assessment paradigm of the last 15 years. The power of a standards-based accountability framework lies in the public availability of the marking scheme that guarantees consistency and unassailability of the grades (Taras, 2009;

DeLuca, 2012). Measurable outcomes can thus be predicted and controlled, which in turn bolsters the institution's professional status as assessor (Almquvist et al., 2017). A social justice agenda is also operationalised through explicit criteria, as knowledge about assessment is now conceivably accessible to all (Torrance, 2017), which is particularly pertinent in the context of internationalisation of HE. However, as some analysts point out, institutional standards are not always reflective of the priorities of the public but reproduce mandated institutional knowledge (Alderman, 2009; Ashworth et al., 2010).

The entanglement of accessibility issues with public accountability is only part of the complex network of formalised academic assessment and feedback. Taras and Davies (2012) advise that assessors' tacit individual frameworks repeatedly endure over disciplinary norms. Assessment is, in essence, judgement, and this process involves heuristic methods (Brooks, 2012; Crisp, 2013). Tacit knowledge – connoisseurship – plays a major part in judging, but, as Tsoukas (2003) notes, is frequently inarticulable. Shay (2005) warns that judgements may be unreliable, inconsistent and difficult to articulate, but this is not the same as bias or random judgements. It is instead a complex process of 'double reading', in which the interpretative framework of the individual is entangled with implicit disciplinary norms.

The literature is clear in that assessors have personal constructs in mind when assessing a piece of work, and that these holistic constructs are fluid and intuitive (Hunter and Docherty, 2011; Bloxham et al., 2016). Descriptors are often used retrospectively to provide justification, and academic judgement becomes a source of bargaining, a 'shopping around for a grade' across markers (Bloxham et al., 2011). This, as Sadler (2009) points out, ultimately leads to 'indeterminacy' in marker's judgement and can no longer serve the mythos of objectivity. The separation between explicit descriptors and private judgements creates a tension that is meant to be addressed by communities of practice

The collective nature of tacit professional knowledge is located in communities of practice, producing, as Orr (2010) stipulates, a contextualised 'guild knowledge' that builds expertise. Marking schemes serve as structured guidance to these shared understandings, and moderation dialogues help practitioners to develop a common language, in turn elucidating the fuzzy nature of descriptors (Grainger et al., 2008; Adie et al., 2013). However, communities of practice do not automatically share a common understanding, and, as Hudson et al (2017) have found, moderation rarely aids calibration. In addition, as Orr (2007) suggests, moderation itself may draw on extra, uncalibrated and internalised criteria, such as specific characteristics of students. Whilst the effectiveness of the process is not proven, the continued faith in the power of moderation, according to Bloxham and Boyd (2012), lies in an attempt to harmonise the intangible sense of personal and locally agreed standards.

The lack of direct correspondence between the verbalisation of criteria and tacit professional knowledge, as Sadler (2013) proposes, may be due to their linguistic indeterminacy that is unquantifiable. A salient question, asked by Forsyth et al. (2015), is whether assessment literacy is at bottom a linguistic issue. Students from international educational and linguistic backgrounds often struggle with academic concepts, such as analysis, synthesis or critical reflection. The techno-rationalist language of the criteria is further confounded by the fuzziness and malleability of standards. Many commentators highlight how marking schemes are composed of qualifiers, modifiers and hedge words that lack a clear grounding in the qualitative nature of the work. Payne and Brown (2011), for example note how the use of relative and comparative terminology adds vagueness about the accomplishment of a criterion, especially at the threshold level. As Greatorex et al. (2001) indicate, marking schemes are commonly based on intuitive and historical wordings. There is no 'thing-in-itself' to which a description may point, and which may help students to direct their own learning.

The majority of literature acknowledges that assessment has the power to direct students' learning, mainly through the benefits of feedback and feedforward (Sambell et al., 2013; Jessop and Tomas, 2017). Another large sector of the literature proposes that an understanding of marking schemes by the student enables self-regulation, empowerment, and autonomy (Popham, 2011; Price et al., 2012). Students need the transparency of the descriptors, operationalised as feedback, to have an understanding of the quality of their performance. Deeley and Bovill (2017) suggest that an ideal way of enhancing students' assessment literacy is through partnerships between students and assessors. William and Thompson (2008) similarly note that the active involvement of students in their learning through a shared understanding of transparent quality criteria fosters assessment literacy.

Since the language of descriptors, and how it may be repeated in feedback, is frequently considered the main stumbling block to a shared understanding, it seems only logical that ambiguities can be lessened through enabling students to get a 'feel' for a standard expressed in the descriptors. This is only possible, however, if these standards are applied fairly and consistently by the assessors. The 'nested hierarchy' of approaches to assessment literacy proposed by O'Donovan et al. (2008), however, frequently stops short of the 'cultivated' community of practice in which those knowledges are made explicit to students. The key concern in the field is, therefore, how students are often excluded from the tacit judgments of their work, and that tacit judgments by assessors are the rule rather than the exception in assessment situations. Our mixed-method study, therefore, investigates firstly students' attitudes towards level descriptors to see whether they trust that their work is evaluated clearly, and whether they believe are able to act on the feedback being given to them.

Quantitative Study on Level Descriptors

Overview

To investigate students' attitudes towards level descriptors, and whether they assisted in their clearer understanding of the criteria, we conducted a cohort survey in 2016/17 at two time points to see if there was a change in their understanding of descriptors across one academic year, since increasing familiarity with assessment processes can potentially increase their literacy. This involved 99 PGT students at March time point and 39 at June time point. There are three incomplete cases at March time point. This reflects a response rate of 62% ($N = 99$) and 24% ($N = 39$) respectively. The drop in return rates at the June time point is most likely due to 'survey fatigue' at the end of the academic year. Using a 35-item self-administered questionnaire, respondents rated items on a 4-point Likert scale, where 1 = Disagree, 2 = Mainly Disagree, 3 = Mainly Agree, and 4 = Agree. This scale had high reliability; Cronbach's alpha was 0.94 in the March time point and 0.95 in the June time point. Their responses on a 4-point Likert scale also allow us to see if they agree or disagree with each item. The questionnaire items were designed via a systematic literature review to gather information about the connection between *assessment and learning*, *students' confidence in how assessment aligns with the curriculum*, *their confidence in the decisions made based upon the level descriptors*, *their own self-regulation* and *their understanding of level descriptors*. In addition, the respondents were asked to describe their background (home/international student) and familiarity with assessment procedures to suggest a baseline for familiarity. The questionnaire was piloted with the 2015/16 cohort, and changes in wording and the arrangement of the scales were made accordingly. Two cases with missing data are excluded from this analysis using pairwise.

Ethics

This research was conducted according to the University's Research Ethics Procedures and approved by the relevant committee. All students who partook in the research did so on an opt-in basis, following a detailed explanation of the aims and objectives of the research. Students who filled in the questionnaire thus gave their informed consent to participate. The data was initially collected via an online survey, which guaranteed privacy and confidentiality to the participants. Due to poor response rates, this was changed later to a paper copy, and we asked students, if they had not yet participated, to fill in the questionnaire at a programme meeting. This may have exerted some pressures on students to comply with this request, as both researchers were present at that meeting. However, it was clearly explained to all students that participation was voluntary, and that the submission of the questionnaire would not be monitored. The same tactic was used at June time point. Since the response rate remained relatively low in relation to the students present at the meetings, it can be assumed that students freely exercised their right not to fill in the questionnaire. The questionnaire did not collect any identifying personal data beyond some general information about knowledge of assessment procedures and whether they were home or international students.

Factor analysis

Exploratory factor analysis (EFA) was used for the March time point dataset, but not in June, since the number of participants was much smaller. EFA sought to explain a larger set of variables with a smaller set of latent constructs, and to determine if the dataset can be reduced to a smaller set of factors (Field, 2013; Hair et al., 2010; Henson and Roberts, 2006). For conducting EFA, a Mahalanobis Distance (MD) for each case was computed to identify multivariate outliers (Hair et al., 2010). The critical value of χ^2 (35) = 66.62 and $\alpha = 0.001$ of df = 35 was taken as the critical value. The result shows that there were no multivariate outliers among the cases. Moreover, distributions of the 35 variables (based on the

questionnaire items) were examined with the frequencies. Although the sample size was small, each of the variables had skewness or kurtosis within acceptable ranges, ±1.

Since the sample size of the survey was only 99 cases, an approach for factor analysis with small sample numbers designed by Zhao (2009) was adopted. Kaiser-Mayer-Olkin (KMO) measure and Bartlett's test were used to check the factorability and sampling. The overall KMO was 0.831 and Bartlett's test was statistically significant ($\chi^2 = 1387.6$, df = 91, p = 0.000), indicating that the sample size was adequate ($p < 0.001$). Furthermore, all the individual variables had an anti-image correlation matrix of less than 0.60, which revealed sample adequacy.

Principal component analysis, using both orthogonal and oblique rotations, was used on all 35 items. Items with the smallest communality were dropped in the analysis until the communalities of all variables were above 0.60. On this basis, 2 items were removed. The mean value of the communalities of 33 items was 0.72 (> 0.70). The scree plot test was applied to determine the number of factors and suggested that a 3-factor solution should be appropriate. The current study set the cut-off point of 0.55 and above for each factor loading suggested by MacCallumet et al. (2001). As a result, 22 items in ***Table 1*** were retained, and there is no cross loading among the 3 factors. There were very weak or negligible correlation between the factors (Factor 1/Factor 2, r = -0.060; Factor 1/Factor 3, r = 0.482; Factor 2/Factor 3, r = -0.006). The variable to factor ratio is 7.3. According to Zhao (2009), this can be regarded as "a moderate to high degree of overdetermination".

Table 1. Factor loadings for PGT students from Explanatory Factor Analysis using Varimax method (N = 97)

Variable	Factor Loadings		
	Conf.	**Value**	**Conc.**
Q5-3. I clearly understand what is meant by a particular grade based on the standard guidelines available through the level descriptors.	0.851		
Q5-4. I can identify with the statements of achievement in level descriptors that are composed with the help of qualifiers, modifiers and hedge words.	0.824		
Q5-2. I find the qualifying words used in the level descriptors helpful for distinguishing grades.	0.807		
Q5-1. I clearly understand what is 'good' or 'poor' achievement of a criterion based on the level descriptors.	0.790		
Q5-5. I find level descriptors can equally be used for any written assignment that is required on the programme (does not apply to Research Methods).	0.726		
Q5-6. I find level descriptors provide fixed reference points of how the criterion has been achieved.	0.722		
Q6-8. I find that the level descriptors make me more satisfied with the marking process.	0.709		
Q6-2. I have a sense of empowerment and autonomy, because the level descriptors provide a clarified expectation of what I need to do in order to improve.	0.703		
Q6-5. I find that level descriptors provide me with a 'feel' for a standard and how standards are applied fairly and consistently.	0.700		
Q6-6. I find that level descriptors increase my confidence in the marking process.	0.681		
Q6-4. I find that the level descriptors help me understand what is behind higher-order skills, such as analysis, synthesis and critical reflection.	0.654		
Q6-7. I find that level descriptors enable me to manage my expectations about the marking process.	0.624		
Q2-1. The level descriptors are explicitly linked to the learning outcomes of the courses on the programmes.		0.784	
Q3-4. The level descriptors underpin the relationships between assessment, learning outcomes and course objectives.		0.691	
Q2-2. Each level descriptor relates to a discrete level of intellectual performance with which I am familiar.		0.664	
Q3-5. The overall quality of my work shows in terms of the multiple interconnected level descriptors for the criteria.		0.650	
Q2-5. The level descriptors refer to the mandated knowledge I have acquired in the courses on the programme.		0.642	

Variable	Factor Loadings		
	Conf.	**Value**	**Conc.**
Q4-2. Assessors may sometimes use more constructs, or rank constructs differently or interpret shared constructs differently, than are stated in the level descriptors.			0.827
Q4-1. Assessors may sometimes have different expectations and relative standards that are not specified in the level descriptors.			0.759
Q4-7. Level descriptors do include a 'hidden curriculum', i.e. interpretations of constructs that are invisible to me.			0.756
Q4-4. Assessors may use 'guild knowledge' (Orr 2010), i.e. professional knowledge that is situated and local, which differs from my own knowledge about the assessment.			0.713
Q4-5. Level descriptors refer to slippery and opaque concepts that can only be known through experience and training.			0.582
Eigen values	8.85	2.83	1.55
% of variance accounted for	40.23	12.84	7.02
Cronbach's α	0.94	0.82	0.78

Findings

In total, the 3-factors accounted for nearly 60 percent of the total variance in the dataset. The 3 factors were labelled "confidence" (12 items), "value" (5 items), and "concern" (5 items). "Confidence" relates to the language of the level descriptors (Q5), and how confident students are in the decisions made based upon these descriptors (Q6). "Value" relates to whether students believed that feedback based on the level descriptors can direct learning and connects with learning outcomes (Q 2 and 3). "Concern" relates to items in which students voiced a belief that assessors drew on tacit and guild knowledges and hidden curricula (***Table 1***). The mean scores for each factor ranged from a minimum value of 1.50 to a maximum value of 4.0. The mean scores of respondents for each of the 3 factors are presented in ***Table 2*** and ***Table 3***. Overall, the respondents had the most positive attitude towards value (range = 1.50–4.00, mean = 3.11, SD = 0.52), but less positive attitude towards

confidence (range = 1.50–4.00, mean = 2.96, SD = 0.60). However, respondents also agreed with the dimension of the factor concern (mean = 2.96, SD = 0.60), indicating that students' felt themselves excluded from that tacit knowledges of the assessors.

Table 2. Means, standard deviation and P-value for two independent samples (Nationality) at March-time point (SD in parentheses)

	Nationality		
	Home students (N = 11)	International students (N = 85)	**P-value**
Confidence	2.86 (0.62)	2.97 (0.61)	0.575
Value	2.98 (0.55)	3.13 (0.52)	0.392
Concern	2.95 (0.71)	3.04 (0.47)	0.691

Table 3. Means, standard deviation and P-value for two independent samples (Familiarity level) at March-time point (SD in parentheses)

	Familiarity level		
	Students who were familiar (N = 47)	Students who were not familiar (N = 49)	**P-value**
Confidence	3.19 (0.53)	2.73 (0.60)	0.000
Value	3.23 (0.47)	2.99 (0.55)	0.026
Concern	2.95 (0.50)	3.09 (0.50)	0.171

This three-factor model was applied to the June time point data to compare the differences between March and June time points, and to see if increased familiarity may change the overall tendency to be concerned about tacit knowledges.

Independent sample T-test

An independent sample T-test analysis was conducted to analyse the difference of the means between home and international students, and students who were familiar and those

who were unfamiliar with the assessment used on the MSc.

March time point: As can be seen in ***Table 2,*** there were no significant differences in mean scores on any of the three factors between home and international students. However, significant differences were found between students who are familiar and ones not familiar with the assessment procedures on confidence ($p < 0.001$) and value ($p < 0.05$) (see ***Table 3***). Independent sample t-test revealed that students, who reported that they are familiar with the assessment as used on the programme, had higher scores on the confidence subscales (mean 3.19, SD = 0.53) and value subscales (mean = 3.23, SD = 0.47) than those who are not familiar with them (mean = 2.73, SD = 0.60, p = 0.000; mean = 2.99, SD = 0.55, p = 0.026 for confidence and value respectively), thus indicating the value of familiarity for assessment literacy.

Comparing March and June time points: Compared to the March group, the majority (32 out of 39) in the June group reported that they had not been familiar with the assessment when they started the programme. Students in the June group reported higher mean scores on "Confidence" (mean = 2.97 versus mean = 2.73) and "Value" (mean = 3.11 versus mean = 2.98), but not on "Concern" (mean = 3.01 versus mean = 3.09).

Table 4. Means, standard deviation and P-value for two independent samples who were familiar with level descriptors at March-time and June-time points (SD in parentheses)

	Students who were familiar		
	March (N = 47)	June (N = 7)	P-value
Confidence	3.19 (0.53)	3.12 (0.45)	0.750
Value	3.23 (0.47)	3.26 (0.40)	0.884
Concern	2.95 (0.50)	3.37 (0.50)	0.042

The results of independent sample T-test summarised in ***Table 4*** implies that the June-group students, who reported that they were familiar with the marking procedures and level descriptors used on the MSc, had higher scores on the confidence subscales (mean = 3.37, SD = 0.50) than the ones in the March group (mean = 2.95, SD = 0.50).There were no significant differences in any of the three-factor scores between participants who were not familiar with level descriptors at both March and June time points (***Table 5***).

Table 5. Means, standard deviation and P-value for two independent samples who were not familiar with level descriptors at March-time and June-time points (SD in parentheses)

	Students who were not familiar		
	March (N = 49)	June (N = 32)	P-value
Confidence	2.73 (0.60)	2.97 (0.53)	0.114
Value	2.99 (0.55)	3.18 (0.49)	0.175
Concern	3.09 (0.50)	3.01 (0.60)	0.172

At March time point, the level of familiarity with level descriptors is related to confidence, value and concern scores. Students with high familiarity reported higher scores on confidence and value, but lower on concern. However, higher levels of the subscale of concern were reported by those with high familiarity in the post-group. This suggests that increasing familiarity with assessment procedures does not alleviate any concerns about tacit knowledges, but may, in fact, increase them.

In summary, concern is thus the only factor that shifted significantly, and in a direction that could be interpreted as negative, and is loosely correlated to the increasing familiarity with assessment procedures.

Qualitative Study

The role of practitioners' beliefs for this topic is considered important, as it lays open the extent to which standards have been internalised by scorers, and its effect on scoring

practices (Bloxam and Boyd, 2012). We decided, therefore, to add a qualitative investigation into practitioners' belief to probe this angle of the topic.

Focus group interview

Four scorers who mark on the programme participated in the focus group interview. Two of the scorers were experienced full-time staff, whereas the other two were final year PhD students who were new to the marking process. The aim of the interview was to have participants share their experience and opinions on how they use the descriptors to assess assignments and see how much level descriptors help with a shared understanding between assessors and, indirectly, students.

The focus group interview was recorded, and notes were taken simultaneously to capture participants' responses accurately. The recording was professionally transcribed and thoroughly read. The data was analysed according to Gale's (2013) 7 stages of Framework Method which is relevant for thematic analysis of interview data. The systematic approach is useful for data to be compared and different perspectives to arise while closely reflecting on the contexts from which they emerge. The data was coded independently and interpreted by two research team members. The emerging themes were highlighted and developed both from the strands in the literature pertaining to research on assessment criteria in HE, and from the narratives of the participants.

Ethics

The potential ethical risk to the research lies in the familiarity of all scorers with each other and with the researchers. Participants may have felt cautious about communicating truthfully how they used level descriptors when marking assignments. The junior participants may have felt pressured to conform to expectations by the more senior staff, and the researcher

who mediated the focus group. Such ethical concerns were mitigated by creating a friendly and collegiate atmosphere during the focus group. Participants were invited to respond to each other and create their own dynamic in the focus group, which would alleviate any interviewer bias (Browne, 2016). As guaranteeing anonymity was problematic, it was decided that the reporting of the data from the focus group research was not attributed to any identifiable variable, such as experience or gendered pronouns.

Findings

Six themes emerged from the data: accountability, rigour, interpretation, language and familiarity, interpretation.

Accountability: Some participants expressed concerns that some scorers tend to give marks based on holistic impressions, which makes it difficult to account for the marks awarded and may contribute to unfairness. They also felt there is a gap in the descriptors in the form of specific features to distinguish the upper and lower bands in the assessment criteria (i.e discriminate clearly between e.g 58% (upper C) and 61% (lower B)). Thus, questions remain as to how scorers account for high or low marks in a band. Despite this, the participants agree the descriptors provide a means to justify the scores and the feedback, and that having the descriptors mean they are able to dispel any uncertainties they have over the essays.

Rigour: Participants felt that scoring becomes efficient and robust only when the descriptors are well understood and have been internalised. They reported slow progress with marking, as they had to take some time to study the language and reflect on their meanings. They felt they had to rationalise the descriptors to understand the allocation of marks. One participant felt that in working with the descriptors, they needed to establish their relationship with the assignment requirement, which they did by matching the different aspects of the

assignment to the assessment criteria. It is clear from here that familiarity of the descriptors paved the way for more informed judgement.

Despite this, some reported falling back on the use of common sense and general impression rather than using the descriptors as a guide. Even when the descriptors for each band are prescribed, scorers tend to assess based on their knowledge, expertise and working standard. One participant expressed that having deep understanding of the course and *"…knowing what the students should actually portray in the essay…"* allows them to have an impression of how an A paper should look like. Others take a more rationalist view on scoring in that they use the presence or absence of a criterion to make sense of their own grading system *"…so if all the suggested changes have no consideration, that's a low B or low C…if there is some consideration then it actually would be a borderline"*.

Language: Most participants felt the language of the descriptors played a role in the accessibility of the assessment criteria. They reported having to work with quite a lot of information within each criterion, at times with language they considered very academic, vague and verbose. They felt the descriptors could be more intelligible to scorers and international students. Despite this, two of the participants expressed appreciation for the more specific phrases in the descriptors, as they were able to match them against their own assessment requirement; the examples provided were helpful to illustrate specific criteria and contribute to better understanding. These factors have affected the identification of criteria and how well scorers use the descriptors.

Familiarity: The participants reported having varying levels of familiarity with descriptors. One participant in particular felt that their lack of familiarity made it difficult to work with them, as it was their first time as a scorer. Another participant stated that having worked with and internalised previous descriptors made it harder to adapt to the existing ones *"…I had internalised these, I was familiar with them, nobody likes change…"*. Others, who

considered themselves familiar with the descriptors, were able to match them quite quickly against the course requirement, and one even reported using the criteria for the purpose of students' self-assessment in their teaching *"...I try to scaffold the students' use of the learning descriptors, so I have like self-assessment checklists…"*. One participant stressed the importance of knowledge and awareness of the descriptors, and the need to be trained in order to be familiar with the assessment process.

Interpretation: The data highlight concerns over different levels of interpretation of descriptors, especially during the standardisation and moderation process. The participants felt this may be because the same generic criteria are used for different courses with varying focus and requirement. One participant felt the more pressing issue is not how the descriptors are used, rather *"...how you interpret the criteria into the context of your course…"*. It was believed that the lack of common understanding of the descriptors has resulted in inconsistencies in marks awarded. One participant expressed their shock over the level of subjectivity surrounding the understanding of the descriptors, which led to differing scores awarded for one single assignment *"...I was shocked by the way some people would give the same essay 70 and some people would give 60 and some people would give 50, and some would give like 48…so there will always be subjectivity…"*.

Participants reported using various strategies for making sense of the descriptors to inform their judgment. Most times, interpretation is subjective *"...I work my way out actually to how to try to understand them…"*. One person said they focused on one criterion at a time and identified a defining phrase and feature for each criteria and band by paraphrasing the descriptors *"...if this essay has A and B, then it is within the knowledge and understanding it belongs to category A…"*. Another participant reported they rely on personal judgement *"…that's where the personal judgement does come in, because I did use a lot of it and it made so much easier…"* and common sense *"...I use common sense, just common sense and I know*

that it must have sounded bad…but I think it's important to me…". Other techniques include relying on examples to interpret meanings, cross referring the new descriptors to the ones used previously, formulating their own understanding of the descriptors based on own topic knowledge and previous marking experience, and also matching descriptors against assignment context.

When there is a difference across marks given, a participant suggested scorers could unpack the descriptors and discuss each other's understanding in order to reach an agreement. However, it was unanimously agreed that building a community of practice amongst scorers requires time, willingness and commitment, which was not always forthcoming.

On the other hand, some participants felt the standardisation and moderation procedure is rigorous, which enabled scorers to go through a norming process. It is an avenue where scorers can discuss the descriptors according to the requirement of the course assignment and come to a common understanding *"...you need generic criteria, but the way people kind of interpret them, that makes the difference."*

Discussion

The qualitative data suggest the descriptors enhanced scorers understanding of the criteria to a certain extent, but the assessment process is fraught with personal issues due to the different ways markers view their own professional knowledge, their topic knowledge and the level they are working at, and the relationship they have with other scorers. The data also show that scorers tend to draw on their own expertise when marking assignments, by formulating their own interpretations of the descriptors, as well as relying on personal judgement and common sense (Bloxham et al., 2016; Shay, 2005). This seems to suggest that while standards have been internalised by scorers, they also resort to tacit judgements that are not made explicit.

This tallies with students' concern over the use of knowledge and criteria in assessment that are not reflected in the descriptors.

The data from both quantitative and qualitative analysis hint, therefore, at a distance between students and scorers in matters of assessment. Level descriptors, as part of the marking scheme, are generally attributed an important role in alleviating student dissatisfaction with assessment and feedback, as they can tackle students' lack of clarity about assessment requirements. In short, level descriptors can enhance assessment literacy. However, the mere existence of descriptors is not enough, since it is increased engagement with them, e.g. via self- and peer assessment, that serves as a training ground for assessment literacy (Bloxham and West, 2004; Rust et al., 2005; Blair and McGinty, 2013; Mulder et al., 2014). Our findings suggest that the distance in assessment practices must be overcome and scorers and students need to become partners in assessment (Smith et al., 2013; Deeley and Bovill, 2017). This is especially important for international students who may not be familiar with assessment concepts. Students and scorers should be constantly engaged in a dialogue about the practices of assessment and the interpretation of criteria, as well as the language used. Exemplars that are shared between scorers and students may encourage the development of connoisseurship of both partners (Handley and Williams, 2011).

As our research indicated, the main hindrance to this ideal is the perceived and acknowledged existence of tacit professional knowledges. The implication is that the use of descriptors alone is no guarantee to facilitate commonality in understanding between students and scorers. In fact, descriptors have emerged as a space where common understanding between assessors and students diverge rather than converge. However, only if students are aware of what is expected of them, and scorers are transparent and accountable in their assessment, can commonality of understanding reasonably be achieved.

To enhance assessment literacy amongst scorers and students, they need to have ownership of the marking scheme they are using. In that sense assessment tasks need to be negotiable and contextual. Familiarity with assessment procedures needs to be honed based on the socialisation of students, especially from other academic cultures, into the system. Feedback should draw on students' multicompetences to analyse, discuss and apply assessment criteria to work, e.g. via dialogic reflection. Assessment literacy is an iterative process, which depends on unhurried chances to develop complex understandings. It is the necessity of constant active engagement with assessment practices to foster assessment literacy. Future directions in that field must, therefore, explore how this can be achieved against increasing demands on academics to assess and provide feedback with ever shorter resources.

Limitations

This research was originally conducted as a pilot validation study for newly-designed level descriptors on the programme. The robust analysis of the literature and the data, however, provide some assurances of the external validity of the research that goes beyond a mere validation of the artefact, and allows a critical analysis of assessment literacy in higher education. As a small-scale project, it does not claim generalisability. It does, however, confirm the salient themes discussed in the literature. It encourages further investigation of professional judgement in assessment, and in how far it may be possible to make international students part of these decisions. Whether level descriptors in criterion-based marking schemes ultimately provide the right pathway is, however, questioned.

Acknowledgements:

We are grateful for the constructive comments of the editor and reviewers of this paper. We thank the University of Edinburgh IAD for their support in terms of the PTAS grant to carry out this research.

References:

Adie, L., Lloyd, M., & D. Beutel. 2013. "Identifying discourses of moderation in higher education." *Assessment & Evaluation in Higher Education* 38 (8): 968–977.

Alderman, G. 2009. "Defining and measuring academic standards: A British perspective." *Higher Education Management and Policy* 21 (3): 9–22.

Almquvist, C. F., Vinge, J., Väkevä, L., & O. Zandén. 2017. "Assessment *as* learning in music education: The risk of "criteria compliance" replacing "learning" in the Scandinavian countries." *Research Studies in Music Education* 39 (1): 3–18.

Ashworth, M., Bloxham, S., & L. Pearce. 2010. "Examining the tension between academic standards and inclusion for disabled students: the impact on marking of individual academics' frameworks for assessment." *Studies in Higher Education*, 35 (2): 209–223

Balla, J., & P. Boyle. 1994. "Assessment of student performance: a framework for improving practice." *Assessment & Evaluation in Higher Education*, 19 (1): 17–28.

Biggs, J. B., & Tang, C. 2010. "Applying constructive alignment to outcomes - based teaching and learning." *Retrieved 23 September 2017 from* https://intranet.tudelft.nl/fileadmin/Files/medewerkersportal/TBM/Onderwijsdag_2014/What-is-ConstructiveAlignment.pdf.

Blair, A., & S. McGinty. 2013. "Feedback-dialogues: Exploring the student perspective." *Assessment & Evaluation in Higher Education* 38 (4): 466–476.

Bloxham, S. 2009. "Marking and moderation in the UK: false assumptions and wasted resources." *Assessment & Evaluation in Higher Education* 34 (2): 209–220

Bloxham, S., & P. Boyd. 2012. "Accountability in grading student work: securing academic standards in a twenty-first century quality assurance context." *British Educational Research Journal* 38 (4): 615–634.

Bloxham, S., den Outer, B., Hudson, J., & M. Price. 2016. “Let’s stop the pretence of consistent marking: exploring the multiple limitations of assessment criteria.” *Assessment & Evaluation in Higher Education* 41 (3): 466–481.

Bloxham, S., and A. West. 2004. “Understanding the Rules of the Game: Peer Assessment as a Medium for Developing Students’ Conceptions of Assessment.” *Assessment & Evaluation in Higher Education* 29 (6): 721–733.

Bloxham, S., & P. Boyd. 2012. “Accountability in grading student work: Securing academic standards in a twenty-first century accountability context.” *British Educational Research Journal* 38 (4): 615–634.

Bloxham, S., Boyd, P., & S. Orr. 2011. “Mark my words: the role of assessment criteria in UK higher education grading practices.” *Studies in Higher Education* 36 (6): 655–670.

Boud, D. 2007. “Reframing Assessment as if Learning were Important.” In D. Boud & N. Falchikov (eds). *Rethinking Assessment in Higher Education* (pp.181–197). London: Routledge.

Brooks, V. 2012. “Marking as judgment.” *Research Papers in Education* 27 (1): 63–80.

Brown, R. 2010. “The current brouhaha about standards in England.” *Quality in Higher Education* 16 (2): 129–137.

Browne, A. L. 2016. “Can people talk together about their practices? Focus groups, humour and the sensitive dynamics of everyday life.” *Area* 48 (2): 198–205.

Crisp, V. 2013. “Criteria, comparison and past experiences: How do teachers make judgements when marking coursework?” *Assessment in Education: Principles, Policy and Practice* 20 (1): 127–144.

Deeley, S. J., & C. Bovill. 2017. “Staff student partnership in assessment: enhancing assessment literacy through democratic practices.” *Assessment & Evaluation in Higher Education* 42 (3): 463–477.

DeLuca, C. 2012. “Preparing teachers for the age of accountability: Toward a framework for assessment education.” *Action in Teacher Education* 34 (5-6): 576–591.

Field, A. 2013. *Discovering Statistics using IBM SPSS Statistics* (4th ed.). London: Sage.

Forsyth, R., Cullen, R., Ringan, N., & M. Stubbs. 2015. “Supporting the development of assessment literacy of staff through institutional process change.” *London Review of Education* 13 (3): 34 – 41.

Gale, K. N., Heath, G., Cameron, E., Rashid, S., & S. Redwood. 2013. “Using the framework method for the analysis of qualitative data in multi-disciplinary health research.” *BMC Medical Research Methodology* 13 (117): 1–8.

Glaser, R. 1963. “Instructional technology and the measurement of learning outcomes: some questions.” *American Psychologist* 18 (8): 519–521.

Grainger, P., Purnell, K., & R. Zipf. 2008. “Judging quality through substantive conversations between markers.” *Assessment & Evaluation in Higher Education* 33 (2): 133–142.

Greatorex, J., Johnson, C., & K. Frame. 2001. “Making the grade: Developing grade descriptors for accounting using a discriminator model of performance.” *Westminster Studies in Education* 24 (2): 167–181.

Hair, J.F., Anderson, R.E., Tatham, R.L. & W. C. Black. 2010. *Multivariate Data Analysis* (7th ed.). New Jersey: Prentice-Hall.

Handley, K., & L. Williams. 2011. “From copying to learning: Using exemplars to engage with assessment criteria and feedback.” *Assessment & Evaluation in Higher Education* 36 (1): 95–108.

Henson, R. K., & J. K. Roberts. 2006. “Use of exploratory factor analysis in published research.” *Educational and Psychological Measurement* 66 (3): 393–416.

Hudson, J., Bloxham, S., den Outer, B., & M. Price. 2017. “Conceptual acrobatics: Talking about assessment standards in the transparency era.” *Studies in Higher Education* 42 (7): 1309–1323.

Hunter, K., & P. Docherty. 2011. “Reducing variation in the assessment of student writing.” *Assessment & Evaluation in Higher Education*, 36 (1): 109–124.

Hussey, T. and P. Smith. 2002. “The trouble with learning outcomes.” *Active Learning in Higher Education* 3 (3): 220–233.

Jessop, T., & C. Tomas. 2017. “The implications of programme assessment patterns for student learning.” *Assessment & Evaluation in Higher Education* 42 (6): 990–999.

Knight, P. 2006. “The local practices of assessment.” *Assessment & Evaluation in Higher Education* 31 (4): 435–452.

Koh, K. H. 2011. “Improving teachers’ assessment literacy through professional development.” *Teaching Education* 22 (3): 255–276.

MacCallum, R. C., Widaman, K. F., Preacher, K. J., & S. Hong. 2001. “Sample size in factor analysis: The role of model error.” *Multivariate Behavioral Research* 36: 611–637.

Mulder, R. A., Pearce, J. M., & C. Baik. 2014. “Peer review in higher education: Student perceptions before and after participation.” *Active Learning in Higher Education* 15 (2): 157–171.

Nicol, D. 2010. *The Foundation for Graduate Attributes: Developing Self-Regulation through Self and Peer Assessment*. Gloucester: The Quality Assurance Agency for Higher Education.

Nicol, D., Thomson, A., & C. Breslin. 2014. “Rethinking feedback practices in higher education: A peer review perspective.” *Assessment & Evaluation in Higher Education* 39 (1): 102–122.

Norton, L., Norton, B., & L. Shannon. 2013. “Revitalising assessment design: What is holding new lecturers back?” *Higher Education* 66 (2): 233–251.

O'Donovan, B., Price, M., & C. Rust. 2008. "Developing student understanding of assessment standards: A nested hierarchy of approaches." *Teaching in Higher Education* 13 (2): 205–217.

Orr, S. 2007. "Assessment moderation: constructing the marks and constructing the students." *Assessment & Evaluation in Higher Education* 32 (6): 645–656.

Orr, S. 2010. "'We kind of try to merge our own experience with the objectivity of the criteria': The role of connoisseurship and tacit practice in undergraduate fine art assessment." *Art, Design and Communication in Higher Education* 9 (1): 5–19.

Payne, E., & G. Brown. 2011. "Communication and practice with examination criteria. Does this influence performance in examinations?" *Assessment & Evaluation in Higher Education* 36 (6): 619–626.

Popham, W. J. 2011. "Assessment Literacy Overlooked: A Teacher Educator's Confession." *The Teacher Educator* 46 (4): 265–273.

Price, M., C. Rust, B. O'Donovan, K. Handley, and R. Bryant. 2012. *Assessment Literacy*. Oxford: Oxford Brookes University.

PTES 2017. *Postgraduate Taught Experience Survey 2017 - Understanding the experiences and motivations of taught postgraduate researchers*. York: The Higher Education Academy.

The Quality Assurance Agency for Higher Education (QAA). 2015. *The UK Quality Code for Higher Education. Retrieved 24 November 2017 from* http://www.qaa.ac.uk/en/Publications/Documents/quality-code-brief-guide.pdf

Rust, C., O'Donovan, B., & M. Price. 2005. "A social constructivist assessment process model: how the research literature shows us this could be best practice." *Assessment & Evaluation in Higher Education* 30 (3): 231–240.

Sadler, D. R. 2009. "Indeterminacy in the use of preset criteria for assessment and grading. *Assessment & Evaluation in Higher Education* 34 (2): 159–179.

Sadler, D. R. 2013. “Assuring academic achievement standards: from moderation to calibration.” *Assessment in Education: Principles, Policy & Practice* 20 (1): 5–19.

Sadler, D. R. 2014. “The Futility of Attempting to Codify Academic Achievement Standards.” *Higher Education* 67 (3): 273–288.

Sambell, K., McDowell, L., & C. Montgomery. 2013. *Assessment for Learning in Higher Education*. London: Routledge.

Shay, S. 2005. “The assessment of complex tasks: a double reading.” *Studies in Higher Education* 30 (6): 663–679.

Smith, C. D., Worsfold, K., Davies, L., Fisher, R., & R. McPhail. 2013. “Assessment literacy and student learning: the case for explicitly developing students’ ‘assessment literacy’.” *Assessment & Evaluation in Higher Education* 38 (1): 44–60.

Stowell, M. 2004. “Equity, Justice and Standards: Assessment Decision Making in Higher Education.” *Assessment & Evaluation in Higher Education* 29 (4): 495–510.

Taras, M. 2009. “Summative assessment: The missing link for formative assessment.” *Journal of Further and Higher Education* 33 (1): 57–69.

Taras, M., & M. S. Davies. 2012. “Perceptions and realities in the functions and processes of assessment.” *Active Learning in Higher Education* 14 (1): 51–61.

Torrance, H. 2017. “Blaming the victim: assessment, examinations, and the responsibilisation of students and teachers in neo-liberal governance.” *Discourse: Studies in the Cultural Politics of Education* 38 (1): 83–96.

Tsoukas, H. 2003. “Do we really understand tacit knowledge?” In M. Easterby-Smith and M. Lyles (eds). *Handbook of Organizational Learning and Knowledge Management*, (pp. 411–427). Cambridge, MA: Blackwell.

William, D., & M. Thompson. 2008. “Integrating assessment with learning: What will it take to make it work?” In C. Dwyer (ed.). *The future of assessment: Shaping teaching and learning* (pp. 53–84). New York: Lawrence Erlbaum Associates.

Zhao, N. 2009. *The minimum sample size in factor analysis*. *Retrieved 27 November 2017 from* https://www.encorewiki.org/display/~nzhao/The+Minimum+Sample+Size+in+Factor+Analysis